Fetch Take home assessment Part 1

Are there any data quality issues present?
Yes, below are the data quality issues encountered during the assessment

- **Duplicate records**
  - Duplicate records in transactions and products dataset. Transactions dataset has 171 duplicate records and Products dataset has 215 duplicate records. User database is clean with respect to duplicates. Below is a check to remove duplicate records and also an example of a duplicate record in the products dataset

```
#removing duplicate records

print("--------Users--------")
print(user_df.shape)
clean_user_df = user_df.drop_duplicates()
print(clean_user_df.shape)

print("--------Transactions--------")
print(trans_df.shape)
clean_trans_df = trans_df.drop_duplicates()
print(clean_trans_df.shape)

print("--------Products--------")
print(prod_df.shape)
clean_prod_df = prod_df.drop_duplicates()
print(clean_prod_df.shape)
```

```
--------Users--------
(100000, 6)
(100000, 6)
--------Transactions--------
(50000, 8)
(49829, 8)
--------Products--------
(845552, 7)
(845337, 7)
```

```
]: prod_df[prod_df['BARCODE']==3498507.0]
```

]:

	CATEGORY_1	CATEGORY_2	CATEGORY_3	CATEGORY_4	MANUFACTURER	BRAND	BARCODE
69016	Snacks	Candy	Gum	NaN	THE HERSHEY COMPANY	ICE BREAKERS	3498507.0
76394	Snacks	Candy	Gum	NaN	THE HERSHEY COMPANY	ICE BREAKERS	3498507.0

- **Null values**
  - There are null values present in all 3 datasets in the below fields. For example, the user dataset has null values in the BIRTH_DATE field. It is important to understand the context of such null values. Presence of these null values can skew the analysis of questions like - **What are the top 5 brands by receipts scanned among users 21 and over? (One of the questions in the assessment)**

Submitted by Raghav Goel

```
# #null values

print("--------Users--------")
print(clean_user_df.isnull().sum())

print("--------Transactions--------")
print(clean_trans_df.isnull().sum())

print("--------Products--------")
print(clean_prod_df.isnull().sum())
```

```
--------Users--------
ID                   0
CREATED_DATE         0
BIRTH_DATE        3675
STATE             4812
LANGUAGE         30508
GENDER            5892
dtype: int64
--------Transactions--------
RECEIPT_ID           0
PURCHASE_DATE        0
SCAN_DATE            0
STORE_NAME           0
USER_ID              0
BARCODE           5735
FINAL_QUANTITY       0
FINAL_SALE           0
dtype: int64
--------Products--------
CATEGORY_1         111
CATEGORY_2        1422
CATEGORY_3       60563
CATEGORY_4      777884
MANUFACTURER    226464
BRAND           226462
BARCODE           3968
dtype: int64
```

- **Inconsistent values**
  - See below screenshot for reference. There is a value called ‘zero’ in FINAL_QUANTITY field of the transactions dataset which could cause errors when trying to convert the column into a numerical format and performing aggregations on it. The value should be ‘0’ instead of ‘zero’ to avoid any data processing issues. Additionally, what does a zero quantity mean?

```
print(list(clean_trans_df['FINAL_QUANTITY'].unique()))
```

```
['1.00', 'zero', '2.00', '3.00', '4.00', '4.55', '2.83', '2.34', '0.46', '7.00', '18.00', '12.00', '5.00', '2.17', '0.23', '8.00', '1.35', '0.09', '2.58', '1.47', '16.00', '0.62', '1.24', '1.40', '0.51', '0.53', '1.69', '6.00', '2.39', '2.60', '10.00', '0.86', '1.54', '1.88', '2.93', '1.28', '0.65', '2.89', '1.44', '2.75', '1.81', '276.00', '0.87', '2.10', '3.33', '2.54', '2.20', '1.93', '1.34', '1.13', '2.19', '0.83', '2.61', '0.28', '1.50', '0.97', '0.24', '1.18', '6.22', '1.22', '1.23', '2.57', '1.07', '2.11', '0.48', '9.00', '3.11', '1.08', '5.53', '1.89', '0.01', '2.18', '1.99', '0.04', '2.25', '1.37', '3.02', '0.35', '0.99', '1.80', '3.24', '0.94', '2.04', '3.69', '0.70', '2.52', '2.27']
```

  - Blank and 0 FINAL_SALE values in transactions dataset. What do these blank and zero sale values mean? Do they represent any kind of discounts to the customer?

```
print(list(sorted(clean_trans_df['FINAL_SALE'].unique())))
```

```
['', '0.00', '0.01', '0.03', '0.04', '0.05', '0.07', '0.09'
 '0.26', '0.28', '0.29', '0.30', '0.32', '0.33', '0.34', '0.35',
```

- Barcode with multiple categories
  - See below an example of a BARCODE in the products database that is associated with two products. This would cause duplication of records when joining with transactions dataset using the BARCODE field. In addition to this, there are also records where a single BARCODE is associated with multiple Brands.

Submitted by Raghav Goel

```
#barcode with two values of CATEGORY_3
clean_prod_df[clean_prod_df['BARCODE']==87108538.0]
```

	CATEGORY_1	CATEGORY_2	CATEGORY_3	CATEGORY_4	MANUFACTURER	BRAND	BARCODE
108325	Snacks	Candy	Mints	NaN	PERFETTI VAN MELLE	MENTOS	87108538.0
300327	Snacks	Candy	Confection Candy	NaN	PERFETTI VAN MELLE	MENTOS	87108538.0

Are there any fields that are challenging to understand?
Below are the fields I think could provide some more context.

- LANGUAGE field in users database -I understand that the language field indicates a language code but if there can be another field called for example ‘Language Name’ that indicates the exact language just to provide any additional layer of information. For example es-419 indicates ‘Latin Spanish’

```
print(list(clean_user_df['LANGUAGE'].unique()))
```

```
['es-419', 'en', nan]
```

- Category fields in products database - The categories are following a hierarchical structure where CATEGORY_4 falls under CATEGORY_3 and CATEGORY_3 falls under CATEGORY_2 and so on. It would be helpful to provide names to the fields like Product Category, Product, SubProduct, etc.

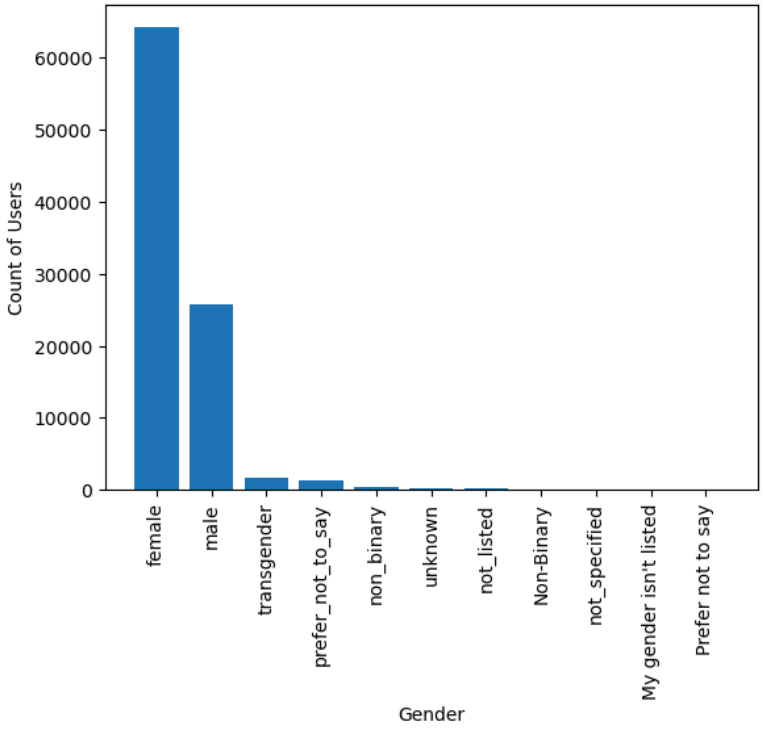
```
#products data

prod_df = pd.read_csv('PRODUCTS_TAKEHOME.csv')

prod_df.head(5)
```

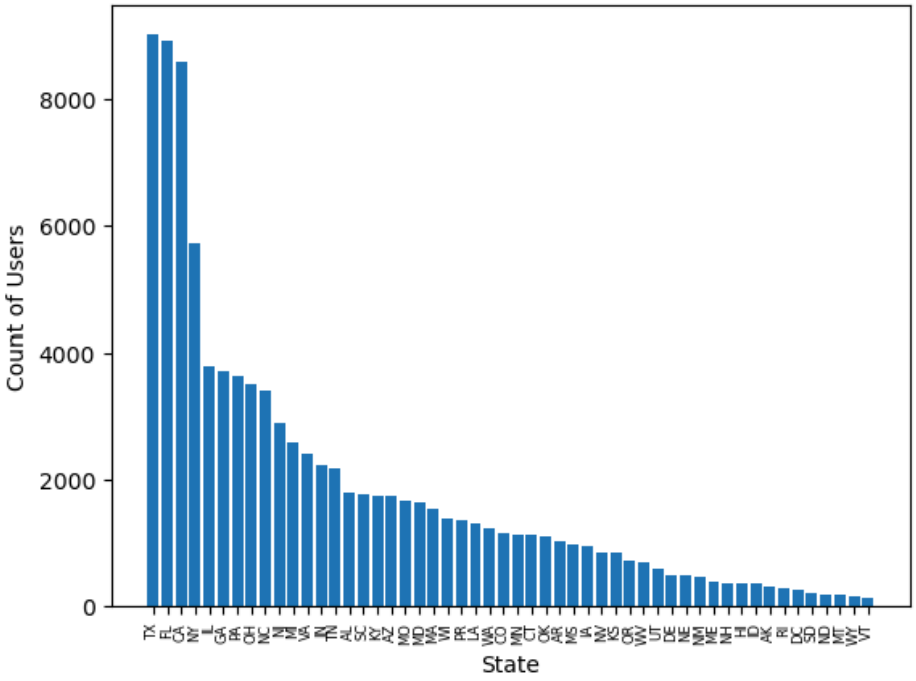
	CATEGORY_1	CATEGORY_2	CATEGORY_3	CATEGORY_4	MANUFACTURER	BRAND	BARCODE
0	Health & Wellness	Sexual Health	Conductivity Gels & Lotions	NaN	NaN	NaN	7.964944e+11
1	Snacks	Puffed Snacks	Cheese Curls & Puffs	NaN	NaN	NaN	2.327801e+10
2	Health & Wellness	Hair Care	Hair Care Accessories	NaN	PLACEHOLDER MANUFACTURER	ELECSOP	4.618178e+11
3	Health & Wellness	Oral Care	Toothpaste	NaN	COLGATE-PALMOLIVE	COLGATE	3.500047e+10
4	Health & Wellness	Medicines & Treatments	Essential Oils	NaN	MAPLE HOLISTICS AND HONEYDEW PRODUCTS INTERCHA...	MAPLE HOLISTICS	8.068109e+11

Submitted by Raghav Goel

Exploratory Analysis using Data Visualization

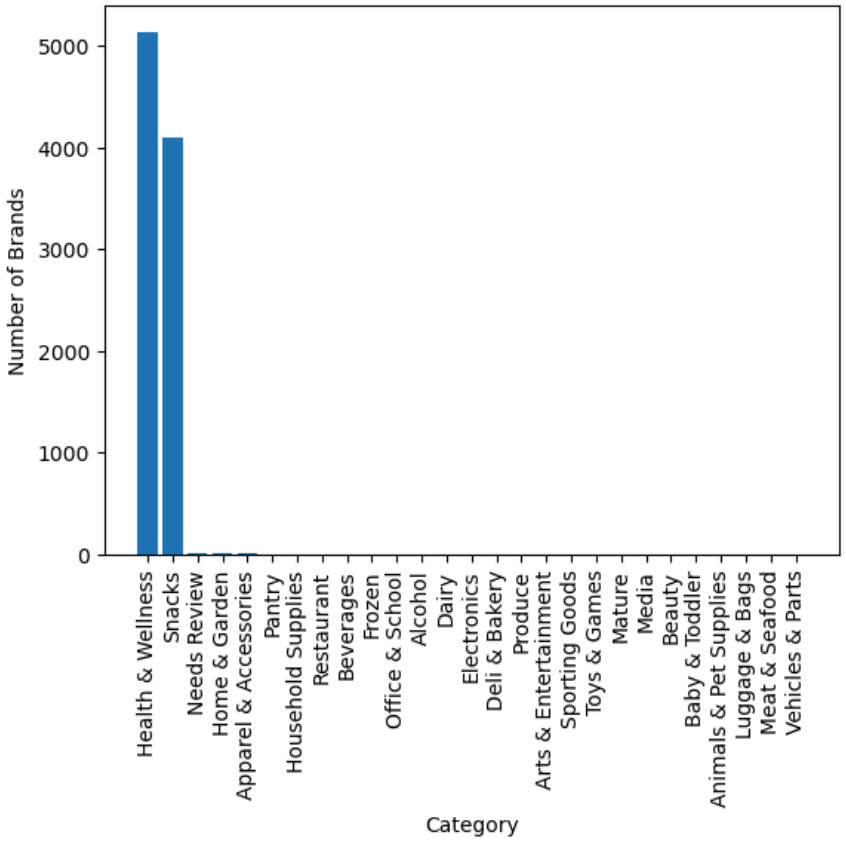
- Number of users
    - By Gender - ~60% of users are Female
    - By State - Most users are from states TX, FL and CA





Submitted by Raghav Goel

- Health & Wellness and Snacks categories have the highest number of brands associated with them. Other Categories do not have many distinct brands.



Submitted by Raghav Goel